KING FAHD UNIVERSITY OF PETROLEUM AND MINERALS

COMPUTER ENGINEERING DEPARTMENT

COE 509 STEGONAGRAPHY ASSIGNMENT PROJECT PROPOSAL

ARABIC DIACRITICS (حركات)-BASED STEGANOGRAPHY

SPRING 2007

TEAM MEMBERS:

MOHAMMED AABED ID# 208701

SAMEH AWAIDEH ID# 260382

ABDUL-RAHMAN ELSHAFEI ID# 260130

INTRODUCTION

This proposal is in response to the Request for Proposals for COE 509 Steganography Assignment, in order to obtain an approval for a topic for Assignment #1. The proposed topic of the project is about utilizing diacritics for Arabic text steganography. The work is a joint collaboration by Mohammed Aabed, Sameh Awaideh and Abdul-Rahman Elshafei. The following sections describes the background of the proposed topic, objectives of the project, approach and implementation, tentative schedule, future work and summarization of related recent research in the area .

BACKGROUND

Steganography, in today's electronic era, is the ability of hiding information in redundant bits of any unremarkable cover media. Information hiding techniques are used to insert extra data into digital media to provide a means for copyright protection (watermarking) or covert communication (steganography). Apart from the numerous methods for embedding information into image, audio and video, schemes for data hiding in text have also been developed. However, very few linguistic steganography schemes were proposed for Arabic text and most of them lacked the capacity for hiding a reasonably amount of data for practical use.

OBJECTIVES AND JUSTIFICATION

The project proposes to introduce a new approach for Arabic text Steganography that can effectively store a large capacity of data while improving stealthiness. The new approach is based on the existence of diacritics in Arabic literals. Diacritic marks in Arabic are used to represent vowel sounds. The literal meaning of Harakāt is "movements", e.g. in the context of moving air waves that we produce while pronouncing vowels. All the consonant sounds

in Arabic are represented by letters but vowel sounds are often not represented in writing. The Harakāt are optional symbols that can be used to represent all the vowels that are not indicated in the ordinary spelling. Data are hidden in Arabic text based on the presence of diacritics. Different types of diacritics would represent either a 0 or a 1 of the hidden data. Having eight major diacritic classes in Arabic (Fatha, Kasra, Damma, Tanwin Fatha, Tanwin Damma, Tanwin Kasra, Sukun and Shadda) preliminary statistics on the subject show that this approach can store large capacity of hidden data compared to the original cover media. It is also noted that this method is a very simple yet novel idea since diacritics will rarely catch the attention of a computer machine or even a normal human reader (unless they are put in the wrong place of course).

METHODS AND TOOLS

Our steganography techniques will be automated; hence we will need a programming language that will be able to do both the encoding and the decoding of the secret message. Our language should be able also to choose a suitable cover media from a list of sources (i.e. HTML documents) and modify this media as needed. It is therefore preferred to have some sort of user friendly (or even GUI) interface. Therefore the proposed tool is either C/C++ or Java.

EXPECTED RESULTS

We expect to produce a program capable of instilling many types of file types into our cover media. The results and comparison phase will try to measure both the capacity and the approximate robustness of the system. Having enough time, we will try to compare the proposed algorithm to similar previously proposed algorithms.

SCHEDULE

- The following table is the project schedule:

Project Component	Date
Researching • Arabic Steganography • Current Text steganography schemes	 Week 1 - 3 Week 1 – 3
Project Proposal	Week 3
Simulation • Coding • Analysis	 Week 4 - 5 Week 4 - 5
Project Report • Writings • Submission	 Week 4 - 5 Week 5
Project Presentation • Preparation • Presentation	 Week 6 - 7 Week 6 - 7

APPENDIX A: SUMARY OF RELATED STEGNAGROPHY ARTICLES

1) INFORMATION HIDING IN TEXT USING TYPESETTING TOOLS WITH STEGO-ENCODING[2]

BY ABDUL-RAHMAN ELSHAFEI ID# 260130

A new steganographic method is proposed by Chao, Shuozhong, and Xinpeng to embed secret information into text files. The method is achieved by making slight modification to scattered inter-word spaces of the formatted text using a typesetting tool called TeX. TeX is particularly useful in generating scientific and technical documents with professional page layout, and unlike MS Word, it does not produce instant formatting results on the screen, but instead, controls the format and page layout via commands and control sequences.

Earlier steganographic methods utilizes TeX program to slightly modify inter-word spaces by using TeX commands to carry secret bits. One approach is to widen/reduce an inter-word space or keep it unchanged to encode the text format with the TeX \hspace command. However, arbitrarily modifying inter-word spaces may affect line-feeding positions. Also if the payload is large, there will be frequent occurrences of such extra spaces that may arouse alert of potential adversaries. To solve this, the method introduced in this paper uses a new scheme such that line ends will not be affected while security is improved. The new scheme states:

- Group each consecutive word pairs together. Words at the end of lines and before or after full stops are ignored.
- Keep spaces between groups unchanged.
- If a 0 is to be embedded, the inter-word space in the group is unchanged.
- If a 1 is to be embedded, the inter-word space in the group is widened for the first encounter, and alternately shrunk and widened for the subsequent 1s. For example, if we add a bit 1, we then increase the inter-word space. If we add another 1 bit, we will do the opposite, which is to decrease the inter-word space. Since the embedded 1s are

alternately widened and shrunk, the total width of inter-word spaces of each line is not significantly changed so that the page layout can be preserved, thus leading to improved stealthiness of the stego-text.

2) LINGUISTC STEGANOGRAPHY: SURVEY, ANALYSIS, AND ROBUSTNESS CONCERNS FOR HIDING INFORMATION IN TEXT [3]

BY SAMEH AWAIDEH ID# 260382

This paper highlights some of the problems inherent in text steganography as well as issues with existing solutions, and describes linguistic problems with character-based, lexical, and syntactic approaches. In addition, the paper explores how a semantic and rhetorical generation approach suggests solutions for creating more believable cover texts, presenting some current and future issues in analysis and generation.

After giving a small introduction to what steganography is, it introduces the reader to steganography main three directions; i.e. image steganography, audio steganography and text steganography. Next, the author gives a summary of the techniques used in image and audio steganography before venturing into the paper's main topic; text steganography.

Text steganography is defined as using "text as the medium in which to hide information". The author claims such a general definition in order to differentiate it from Linguistic steganography which she defines as "linguistically-driven generation and modification of cover texts". After which the author defines the main vocabulary in steganography attack techniques, like what is steganalysis? What is meant by adversarial models? And what are the methods of steganalysis?

In chapter 2, the author divides text steganography into three main categories; format-based methods, random and statistical generation, and linguistic methods. Noting that within each category, the text can either be generated from scratch or embedded within known plaintext. Format based methods are defined as: using the physical formatting of text as a space in which to hide information. Format-based methods generally modify existing text in order to hide the steganographic text. Insertion of spaces or non-displayed characters, deliberate misspellings distributed throughout the text, and resizing of fonts are some of the many format-based methods used in steganography. Whereas random and statistical generation tries to simulate some property of normal text, usually by approximating some arbitrary statistical distribution found in real text to produce its own cover texts. Finally, linguistic steganography specifically considers the linguistic properties

of generated and modified text, and in many cases, uses linguistic structure as the space in which messages are hidden.

Chapter 3 tackles the linguistic concerns with these existing methods. Emphasizing that one level of linguistic correctness is often implied by another; a syntactically correct text will necessarily contain valid lexical items, while a semantically coherent text is generally also syntactically correct, and a rhetorically valid text has coherent semantics and correct grammar. It summarizes that all of the previous methods are attack-prone, especially if a human reader was to analyze the text (which seems to be the trend nowadays in steganalysis).

Finally the fourth chapter gives insight into future directions in constructing linguistically and statistically cover texts. It stresses the importance of considering both coherent semantics and rhetorical structure. It concludes that it is hoped that a system can be devised which can produce semantically, syntactically, lexically, and rhetorically correct cover texts which will be an advance in increasing human readability of cover texts without suspicion, as well as evading statistical and linguistic attacks against such methods.

3) A NEW APPROACH TO PERSIAN/ARABIC TEXT STEGANOGRAPHY [4]

BY SAMEH AWAIDEH ID# 260382

The authors in this paper propose a new method for text steganography which can be categorized under *feature coding methods*. Considering the existence of too many points in Persian and Arabic phrases, in this approach, by vertical displacement of the points, information can be hidden in the texts. Their method has been implemented by JAVA programming language.

The paper starts by stating the difficulties in text steganography over other kinds of steganography. First, the relative lack of redundant information in a text file compared with a picture or a sound file. Second, the structure of text documents is identical with what we observe, while in other types of documents, such as in pictures we can hide information by introducing changes in the structure of the document without making a notable change in the concerned output. However, the authors do note that using text is preferred over other media, because the texts occupy lesser memory, communicate more information and need less cost for printing as well as some other advantages.

Section 2 discusses the previous works done in the field, dividing text steganography into ten categories; Steganography of Information in Random Character and Word Sequences, Steganography of Information in Specific Characters in Words, Creating Spam Texts, Line Shifting, Word Shifting, Symantic Methods, Feature Coding, Abbreviation and Open Spaces. The next section indulges into the authors proposed method.

Section 3 starts by stating that more than half of the Arabic and the Persian characters have points, creating a suitable environment to hide information in them. Before hiding data, they propose to compress the concerned information first. Then, they look for the first pointed letter in the given text. By finding this character, they go to the compressed information and read the first bit of information which has the values of zero or one. If the value of the bit were zero, the concerned character remains unchanged. If the value of the bit were one, they shift the point on the concerned character a little upward. This procedure is repeated for the next pointed characters in the text and the next bits of information. Thus, the entire information is hidden. In order to divert the attention of readers, after hiding all information, the points of the remaining characters are also

changed randomly. Of course, before doing this, the size of hidden information is also hidden in the beginning of the text.

Section 4 discusses the possible advantages and disadvantages of the proposed procedure, whereas Section 5 produces some experimental results. Noting that actual capacity of the authors proposed method is no more than 2%! Finally section 6 concludes that by employing a font editing software, the program can be enabled dynamically to produce necessary fonts for hiding information so that the output form of the text is not homogenous and conform to the input form of the text.

4) A TESTBED FOR EVALUATING SECURITY AND ROBUSTNESS OF STEGANOGRAPHY TECHNIQUES [6]

BY MOHAMMED A. AABED ID# 208701

This work introduces a novel framework for testing different steganography techniques and evaluating the level of security and power in such techniques. The testbed environment works as follows:

- Messages are enclosed into still images
- Preprocessing is performed to obtain the hidden messages
- Post processing stage is applied for further analysis on the image

The proposed environment is designed in a modular orientation that consists of four main components. Here is a description of these four stations [Fig. 1; [6]].

The first component is Steganography toolbox. The function of this component is to create stego-images that contain embedded messages. The production of the stego-images is done using some well known steganography software packages that is contained within the station. Second is the Capturing and preprocessing station. As name of this station shows, this module has two duties. First, it captures stego-images passing through the network and rehabilitates them to their genuine format. Next, the captured stego-image is preprocessed to determine wither or not it has an embedded message. The decision in this stage is formed based on studying few parameters read from the image in its hexadecimal format. Inserted software signature, LSB's replacement by zeros or ones, and inserted spaces blocks are examples of the parameters that might be used. At this step, if an image is detected to be a stego-image, it gets passed to the next station. Otherwise, the image is discarded which accelerated the monitoring process. The third module is the Steganalysis station. This part is responsible for extracting embedded messages in a stego-image using statistical analysis techniques which is based on steganography toolbox defined available in the literature. This toolbox is capable of detecting the deviation in statistical properties from the normal which in turn identifies the hidden message. The fourth and final station is the Distortion Station. This stage utilizes image processing technology to disable the hidden message while keeping the original image intact. If this massage cannot be separated from the cover, then it is ignored by the distortion station.

This approach was applied on a diverse group of image and messages. The tools used in the environment uses four main tool boxes:

- Data Insertion
- LSB (Least Significant Bit)
- Palette manipulation
- DCT (Discrete Cosine Transform) steganography technique

The simulation was done by generating the stego-image using three tools, namely Camouflage, Jpegx and J-Steg. The results showed a very good accuracy percentage for the J-Steg and a very low accuracy percentage for Camouflage and Jpegx. This due to the fact that the detection program, StegDetect, was designed to work with J-Steg, where the added enhancements to the program to work with other tools where defective.

REFERENCES

[1] Wikipedia, "Harakat" [Online document], 2006 Sep. 13, [cited 2007 Mar. 07], Available: http://en.wikipedia.org/wiki/Harakat

[2] Chen Chao, Wang Shuozhong, Zhang Xinpeng: Information Hiding in Text Using Typesetting Tools with Stego-Encoding. ICICIC (1) 2006: 459-462

[3] K. Bennet, "Linguistc Steganography: Survey, Analysis, and Robustness Concerns For Hiding Information In Text", CERIAS Tech Report 2004-13.

[4] M. Hassan Shirali-Shahreza, Mohammad Shirali-Shahreza, "A New Approach to Persian/Arabic Text Steganography", Proceedings of the 5th IEEE/ACIS International Conference on Computer and Information Science and 1st IEEE/ACIS International Workshop on Component-Based Software Engineering, Software Architecture and Reuse (ICIS-COMSAR'06).

[5] N. Johnson, S. Jajodia, "Exploring Steganography: Seeing the Unseen", IEEE Computer, February 1998, vol. 31, no. 2, pp.26-34

[6] Tariq Al –Hawi, MahmoudAI Qutayri, and Hassan Barada, "A Testbed for Evaluating Security and Robustness of Steganography Techniques", *in Proceedings of the 46th IEEE International Midwest Symposium on Circuits and Systems, 2003.* vol. 3, pages 1583- 1586, Dec. 2003.

Instructor Comments:

Your proposal of using diacritic is interesting. However, you need to figure a clear way to retrieve the hidden bits from the stego-file. Also, study how can you hide bits in an already used stego file, it seems that the capacity will go very low. What is the cost if we want to have the same cover file for hiding data with same capacity?

Since you informed me verbally that you have made a statistical study comparing different diacritics and their percentages, I suggest you document the study result that using "fatha" is the best in terms of capacity.

Very interesting idea, hopefully we can get your code running correctly.